A Note on Solving Nearly Triangular Toeplitz Systems

William F. Trench
Department of Mathematics
Trinity University
715 Stadium Drive
San Antonio, Texas 78284

Submitted by Hans Schneider

ABSTRACT

An algorithm is presented which reduces the problem of solving a Toeplitz system (1) $TX = Y$ to simple recursive computations and solving a related Toeplitz system which is of lower order if T is nearly triangular. The method does not require that T or any of its principal submatrices be nonsingular, but only that (1) have a solution X for the given Y. In the case where T^{-1} exists, a formula is given for it in terms of the inverse of an associated matrix discovered by Widom.

1. INTRODUCTION

Here we present a method for solving the linear system

$$TX = Y, \tag{1}$$

where T is the $n \times n$ Toeplitz matrix

$$T = \left(t_{j-i}\right)_{i,j=1}^{n}.$$

There are many known methods for exploiting the simple structure of T so as to solve (1) efficiently. However, we believe that the method presented here is new and that it may be computationally useful in the case where T is "nearly triangular" in a sense which we now formulate.

52 Vanderbilt Ave., New York, NY 10017 0024-3795/87/$3.50

The order n of T is assumed to be fixed throughout the paper. In general, the quantities defined below depend upon n, but we do not make this explicit in the notation. We assume that T is not triangular, since this case is easily handled by known methods. (See Remarks 1 and 2.) Then there are uniquely defined integers r and s such that

$$1 \leqslant r, s \leqslant n-1, \tag{2}$$

$$t_r t_{-s} \neq 0, \quad \text{and} \quad t_m = 0 \quad \text{if} \quad -n+1 \leqslant m < -s \text{ or } r < m \leqslant n-1. \tag{3}$$

Although our results are valid and may be of some theoretical interest even if $r = s = n-1$, the algorithm that we present does not seem to offer any computational advantage unless either r or s is small compared with n, so that T is nearly triangular in an obvious intuitive sense which need not be defined more precisely. (For example, if $r = 1$ or $s = 1$, then T is in Hessenberg or "almost triangular" form [5, p. 147].)

We present our results in a way which is computationally appealing if

$$1 \leqslant r \leqslant s \leqslant n-1, \tag{4}$$

so that T is nearly lower triangular if r is small compared to n; however, we do not make use of the second inequality in (4), so the results are valid under the less specific condition (2). It is easy to adapt the proposed algorithm to the case where T is nearly upper triangular (i.e., $1 \leqslant s \leqslant r \leqslant n-1$, with s small compared with n), since if T is a Toeplitz matrix and J is the matrix obtained by reversing the columns of I, then $J^2 = I$ and $JTJ = T^t$; hence, (1) can be rewritten as $T^t\hat{X} = \hat{Y}$, where $\hat{X} = JX$, $\hat{Y} = JY$, and T^t is nearly lower triangular (and still Toeplitz) if T is nearly upper triangular.

2. THE BASIC THEOREM

Let

$$P(z) = \sum_{m=-s}^{r} t_m z^{r-m},$$

and define $\{u_i\}_{-\infty}^{\infty}$ by

$$[P(z)]^{-1} = \sum_{i=-\infty}^{\infty} u_i z^i,$$

where the series converges for small z; thus,

$$u_i = \begin{cases} 0, & i < 0, \\ t_r^{-1}, & i = 0, \\ -t_r^{-1} \sum_{m=-s}^{r-1} t_m u_{m+i-r}, & i \geqslant 1. \end{cases} \tag{5}$$

[For $i \geqslant 0$, u_i can be expressed explicitly in terms of the zeros of $P(z)$, but we make no use of this here.]

The proposed algorithm stems from the following theorem.

THEOREM 1. *Suppose that* (2) *and* (3) *hold. Let*

$$Y = [y_1, \ldots, y_n]^t$$

be given, and define $v_{-s+1}, \ldots, v_{n+r}$ *by*

$$v_i = \begin{cases} 0, & -s+1 \leqslant i \leqslant r, \\ t_r^{-1}\left[y_{i-r} - \sum_{k=-s}^{r-1} t_k v_{k+i-r}\right], & r+1 \leqslant i \leqslant n+r. \end{cases} \tag{6}$$

Then the vector

$$X = [x_1, x_2, \ldots, x_n]^t \tag{7}$$

satisfies (1) *if and only if*

$$x_i = v_i + \sum_{q=1}^{r} a_q u_{i-q}, \qquad 1 \leqslant i \leqslant n, \tag{8}$$

where the constants $a_1, \ldots, a_r$ *satisfy the* $r \times r$ *Toeplitz system*

$$\sum_{q=1}^{r} a_q u_{n+p-q} = -v_{n+p}, \qquad 1 \leqslant p \leqslant r. \tag{9}$$

Proof. In terms of components, (1) becomes

$$\sum_{j=1}^{n} t_{j-i}x_j = y_i, \qquad 1 \leqslant i \leqslant n. \tag{10}$$

Introducing the new summation index $k = j - i$ enables us to rewrite (10) as

$$\sum_{k=-i+1}^{n-i} t_k x_{k+i} = y_i, \qquad 1 \leqslant i \leqslant n,$$

which becomes

$$\sum_{k=-s}^{r} t_k x_{k+i} = y_i, \qquad 1 \leqslant i \leqslant n, \tag{11}$$

if we define

$$x_i = 0 \qquad \text{for} \quad -s+1 \leqslant i \leqslant 0 \text{ and } n+1 \leqslant i \leqslant n+r. \tag{12}$$

Therefore, the vector X in (7) satisfies (1) if and only if the vector

$$\hat{X} = [x_{-s+1}, \ldots, x_{n+r}]^t \tag{13}$$

satisfies (11) and (12). Now consider a vector $\hat{X}$ as in (13), where

$$x_i = v_i + \sum_{q=1}^{r} a_q u_{i-q}, \qquad -s+1 \leqslant i \leqslant n+r.$$

Then (5) and (6) imply that

$$x_i = 0, \qquad -s+1 \leqslant i \leqslant 0. \tag{14}$$

Moreover, (6) implies that

$$\sum_{k=-s}^{r} t_k v_{k+i} = y_i, \qquad 1 \leqslant i \leqslant n. \tag{15}$$

Also, if $i \geqslant 1$ and $1 \leqslant q \leqslant r$, then $i - q + r \geqslant 1$ and

$$\sum_{k=-s}^{r} t_k u_{k+i-q} = \sum_{k=-s}^{r} t_k u_{k+(i-q+r)-r} = 0, \tag{16}$$

because of (5). This enables us to complete the proof of sufficiency, since (8), (15), and (16) imply (11), while (9) and (14) imply (12).

To prove necessity, suppose that (1) has a solution X as in (7), and let

$$\hat{X} = [0,\ldots,0, x_1,\ldots, x_n, 0,\ldots,0]^t, \tag{17}$$

where there are s leading (in positions $-s+1,\ldots,0$) and r trailing (in positions $n+1,\ldots,n+r$) zeros. Define

$$V = [0,\ldots,0, v_1,\ldots, v_{n+r}]^t$$

[cf. (6)], where there are s leading zeros $v_{-s+1},\ldots, v_0$. Also define

$$Z = \hat{X} - V = [0,\ldots,0, z_1,\ldots, z_{n+r}]^t.$$

Then Z satisfies the system

$$z_i = 0, \qquad -s+1 \leqslant i \leqslant 0,$$

$$\sum_{k=-s}^{r} t_k z_{k+i} = 0, \qquad 1 \leqslant i \leqslant n.$$

[Recall (11) and (15).] This is a system of $n+s$ independent equations in $n+r+s$ unknowns; therefore, its solution space is r-dimensional. It is easily verified that the vectors

$$U_q = [u_{-s+1-q}, u_{-s+2-q},\ldots, u_{n+r-q}]^t, \qquad 1 \leqslant q \leqslant r,$$

form a basis for this space. [Recall (16).] Consequently,

$$\hat{X} - V = Z = \sum_{q=1}^{r} a_q U_q \tag{18}$$

for some choice of constants $a_1,\dots,a_r$, and this implies (8). Moreover, (17) and (18) imply (9). ∎

REMARK 1. If (3) holds with $r = 0$, so that T is lower triangular, then $Y = [v_1,\dots,v_n]^t$ is the solution of (1), with $\{v_i\}$ as defined by (6) with $r = 0$. This is well known, but we point it out for completeness.

3. THE ALGORITHM FOR SOLVING $TX = Y$

Theorem 1 suggests the following algorithm for solving (1):

(i) Compute $u_{-r-s+1},\dots,u_{n+r-1}$ from (5). (This need not be repeated if Y is changed.)

(ii) Compute $v_1,\dots,v_{n+r}$ from (6).

(iii) Solve the $r \times r$ system (9) for $a_1,\dots,a_r$.

(iv) Compute $x_1,\dots,x_n$ from (8).

Accounting for terms which are zero by definition, the number of multiplications required in parts (i), (ii), and (iv) are $(r+s)(2n+r-s-1)/2$, $n+(r+s)(2n-r-s-1)/2$, and $rn - r(r-1)/2$, respectively. The computational cost of part (iii) is determined by r; moreover, since (9) is a Toeplitz system, there are efficient ways for solving it. Therefore, if r is small compared with n, the algorithm requires approximately n^2 multiplications if $s = n-1$. If both r and s are small compared with n, so that T is banded, then approximately $(2r+2s+1)n$ multiplications are required.

Many methods have been devised specifically for solving Toeplitz systems, and some even more specifically for banded Toeplitz systems (e.g., see [1], [2], [4], [9], [10], [12], and [13], for a very incomplete sample). Many of these, such as Bareiss's [1], Dickinson's [4], Zohar's [12, 13], and the author's [9, 10], require that T and all its principal submatrices be nonsingular. Others rely on the fact that if T and its $(n-1)$st-order principal submatrix are both nonsingular, then T^{-1} is completely determined by (and can be efficiently computed from) its first row and column, as was shown by the author in [9]. The algorithm of Brent, Gustavson, and Yun [2] requires only that T be invertible, but the strategy of the algorithm dictates one course of action if the $(n-1)$st principal submatrix is invertible, and another if it is not. Bunch [3] has pointed out that this discontinuity in the algorithm may engender instability if the principal submatrix is nearly singular.

The algorithm presented here requires no assumptions concerning invertibility of the principal submatrices of T, or even of T itself. Theorem 1 implies that if (1) has a solution for a given Y, then so does (9). More

precisely, Theorem 1 implies that there is a one-to-one correspondence between the solutions of (1) and (9) for a given Y. For these reasons, we believe that the algorithm may be a useful additional tool for solving banded Toeplitz systems, for which there are many applications. It should also be pointed out that Sugiyama [8] has recently published an algorithm for solving discrete Wiener-Hopf equations which requires the solution of nearly triangular (but not banded) Toeplitz systems whose matrices may be singular.

4. T^{-1} IN TERMS OF WIDOM'S DETERMINANT

Theorem 1 clearly implies that T is invertible if and only if the $r \times r$ matrix

$$W = (u_{n+p-q})_{p,q=1}^{r}$$

of the system (9) is invertible. This is a known result, established by Widom [11] for Hermitian T and extended to general Toeplitz matrices by Schmidt and Spitzer [7]. Here we assume that $T^{-1} = (\tau_{ij})_{i,j=1}^{n}$ exists, and give a formula for $\{\tau_{ij}\}$ in terms of $\{u_m\}$ [cf. (5)] and the inverse

$$W^{-1} = (\omega_{pq})_{p,q=1}^{r} \tag{19}$$

of Widom's matrix.

THEOREM 2. *Suppose that* (2) *and* (3) *hold, and that T is invertible. Then the elements of T^{-1} are given by*

$$\tau_{ij} = u_{i-j-r} - \sum_{p,q=1}^{r} \omega_{pq} u_{n+q-j-r} u_{i-p}, \qquad 1 \leqslant i, j \leqslant n. \tag{20}$$

Proof. Let j be fixed, $1 \leqslant j \leqslant n$. We apply Theorem 1 to (1) with

$$Y = [\delta_{1j}, \delta_{2j}, \ldots, \delta_{nj}]^t \tag{21}$$

(the jth column of the identity), so that

$$X = [\tau_{1j}, \tau_{2j}, \ldots, \tau_{nj}]^t$$

(the jth column of T^{-1}). With Y as in (21), (6) becomes

$$v_i = \begin{cases} 0, & -s+1 \leqslant i \leqslant r, \\ t_r^{-1}\left[\delta_{i-r,j} - \sum_{k=-s}^{r-1} t_k v_{k+i-r}\right], & r+1 \leqslant i \leqslant n+r, \end{cases}$$

which can be rewritten as

$$v_i = \begin{cases} 0, & -s+1 \leqslant i \leqslant j+r-1, \\ t_r^{-1}, & i = j+r, \\ -t_r^{-1} \sum_{k=-s}^{r-1} t_k v_{k+i-r}, & j+r+1 \leqslant i \leqslant n+r. \end{cases} \tag{22}$$

Comparing (5) and (22) shows that $v_i = u_{i-j-r}$. Therefore, (8) (with $x_i = \tau_{ij}$) becomes

$$\tau_{ij} = u_{i-j-r} + \sum_{p=1}^{r} a_p u_{i-p}, \qquad 1 \leqslant i \leqslant n, \tag{23}$$

and (9) becomes

$$\sum_{q=1}^{r} a_q u_{n+p-q} = -u_{n+p-j-r}, \qquad 1 \leqslant p \leqslant r. \tag{24}$$

From (19) and (24),

$$a_p = -\sum_{q=1}^{r} \omega_{pq} u_{n+q-j-r}, \qquad 1 \leqslant p \leqslant r.$$

This and (23) imply (20). ∎

REMARK 2. If $r = 0$, so that T is lower triangular, then (20) can be interpreted as

$$\tau_{ij} = u_{i-j}, \qquad 1 \leqslant i, j \leqslant n,$$

with $\{u_m\}$ as defined in (5) with $r = 0$. This is a known result; see, e.g., [9].

REMARK 3. Meek [6] has given other formulas for the elements of T^{-1} as ratios of determinants involving $\{u_i\}$. The orders of Meek's determinants also depend upon r, the number of nonzero superdiagonal elements in T.

REFERENCES

1 E. H. Bareiss, Numerical solution of linear equations with Toeplitz and vector Toeplitz matrices, *Numer. Math.* 13:404–424 (1969).

2 R. P. Brent, F. G. Gustavson, and D. Y. Y. Yun, Fast solution of Toeplitz systems of equations and computation of Padé approximants, *J. Algorithms* 1:259–295 (1980).

3 J. R. Bunch, Stability of methods for solving Toeplitz systems of equations, *SIAM J. Sci. Statist. Comput.* 6:349–364 (1985).

4 B. W. Dickinson, Efficient solution of linear equations with banded Toeplitz matrices, *IEEE Trans. Acoust. Speech Signal Process.* ASSP-27:421–423 (1979).

5 A. S. Householder, *The Theory of Matrices in Numerical Analysis*, Dover, New York, 1975.

6 D. S. Meek, The inverses of Toeplitz band matrices, *Linear Algebra Appl.* 49:117–129 (1983).

7 P. Schmidt and F. Spitzer, The Toeplitz matrices of an arbitrary Laurent polynomial, *Math. Scand.* 8:15–38 (1960).

8 Y. Sugiyama, An algorithm for solving discrete-time Wiener-Hopf equations based on Euclid's algorithm, *IEEE Trans. Inform. Theory* IT-32:394–409 (1986).

9 W. F. Trench, An algorithm for the inversion of finite Toeplitz matrices, *J. Soc. Indust. Appl. Math.* 12:515–522 (1964).

10 W. F. Trench, Inversion of Toeplitz band matrices, *Math. Comp.* 28:1089–1095 (1974).

11 H. Widom, On the eigenvalues of certain Hermitian operators, *Trans. Amer. Math. Soc.* 88:491–522 (1958).

12 S. Zohar, Toeplitz matrix inversion: The algorithm of W. F. Trench, *J. Assoc. Comput. Mach.* 16:592–601 (1969).

13 S. Zohar, The solution of a Toeplitz set of linear equations, *J. Assoc. Comput. Mach.* 21:272–276 (1974).

Received 16 January 1986; revised 7 July 1986